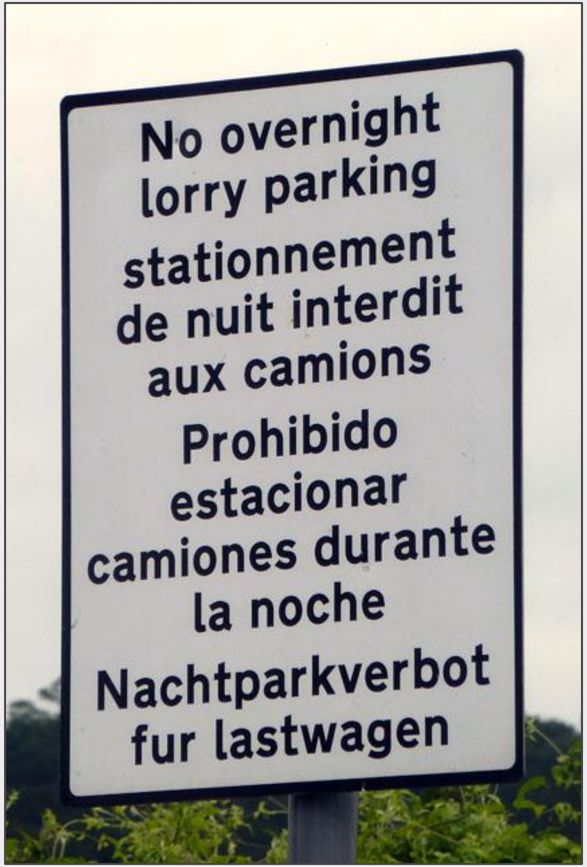
Introduction to Text Analytics

Session 1: May 31, 2018

What is Natural Language Processing?

Methods for computers to parse, interpret, and act on human language data; either speech or text.



NLP Applications

- **Summarization**
- Reference Resolution
- Machine Translation
- Language Generation
- Language Understanding
- **Document Classification**
- Author Identification
- Part of Speech Tagging
- Question Answering
- **Information Extraction**
- Information Retrieval
- Speech Recognition
- Sense Disambiguation
- **Topic Modelling**
- Relationship Detection
- **Named Entity Recognition**

Not* Computational Linguistics

Understanding the formulation and evolution of linguistic symbols: mappings between sight, sound, and mental images.

[The study of grammar], initiated by the Greeks and continued mainly by the French, was based on logic. It lacked a scientific approach and was detached from language itself. Its only aim was to give rules for distinguishing between correct and incorrect forms; it was a normative discipline, far removed from actual observation, and its scope was limited.

-- Ferdinand de Saussure

* (necessarily)

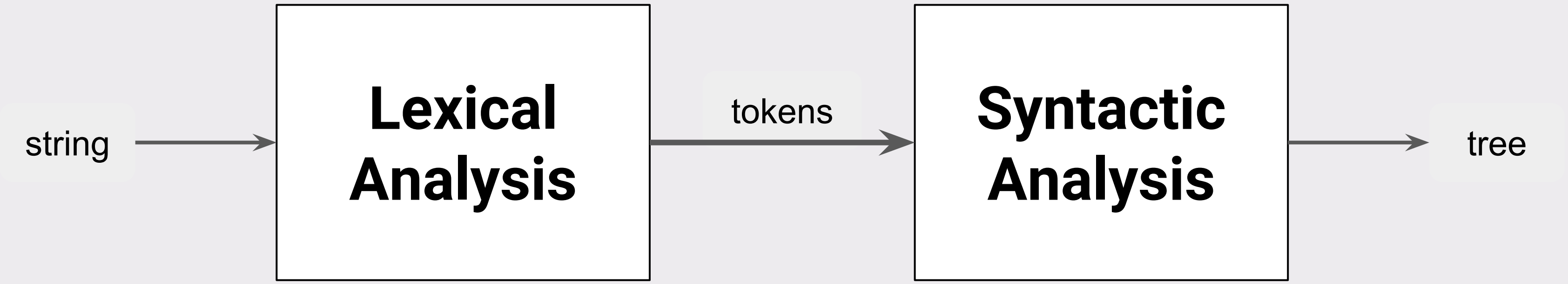
What is language data?

Formal Languages

- Strict, unchanging rules defined by grammars and parsed by regular expressions
- Generally application specific (chemistry, math)
- Literal: exactly what is said is meant.
- No ambiguity
- Parsable by regular expressions
- Inflexible: no new terms or meaning.

Natural Languages

- Flexible, evolving language that occurs naturally in human communication
- Unspecific and used in many domains and applications
- Redundant and verbose; ambiguous
- Expressive
- Difficult to parse
- Very flexible even in narrow contexts



string

Lexical Analysis

tokens

Syntactic Analysis

tree

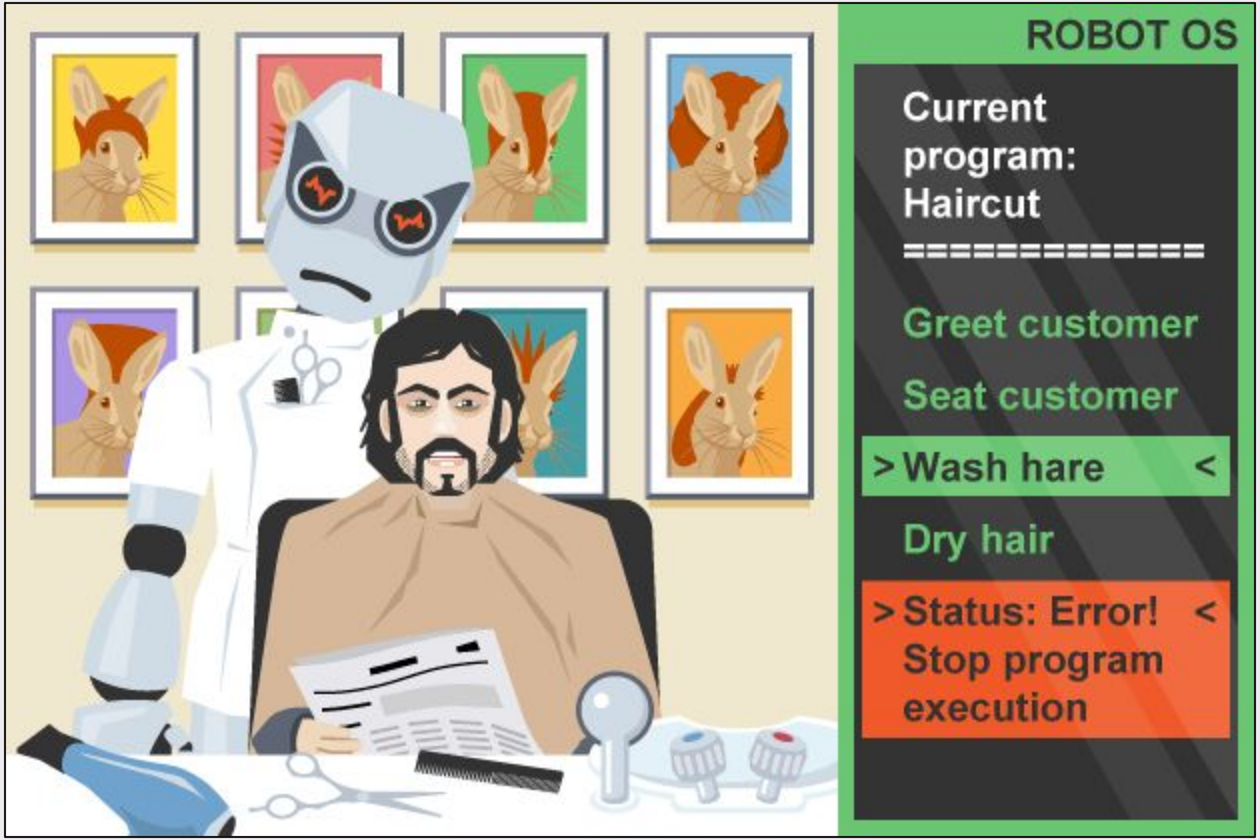
“Lexing”

Regular Grammar
(Word Rules)

“Parsing”

Context Free Grammar
(Sentence Rules)

Computer science has traditionally focused on formal languages



ROBOT OS
Current program: Haircut
==============
Greet customer
Seat customer
> Wash hare <
Dry hair
> Status: Error! <
Stop program execution

Ambiguity is required for understanding when communicating between people with diverse experience.

Challenges in Natural Language Processing

Lexical Ambiguity

“Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo.”

Anaphora Resolution

“John found Jack the love of his life.”

Local Coherence

“The horse raced past the barn fell.”

Structural Ambiguity

“Time flies like an arrow; fruit flies like a banana.”

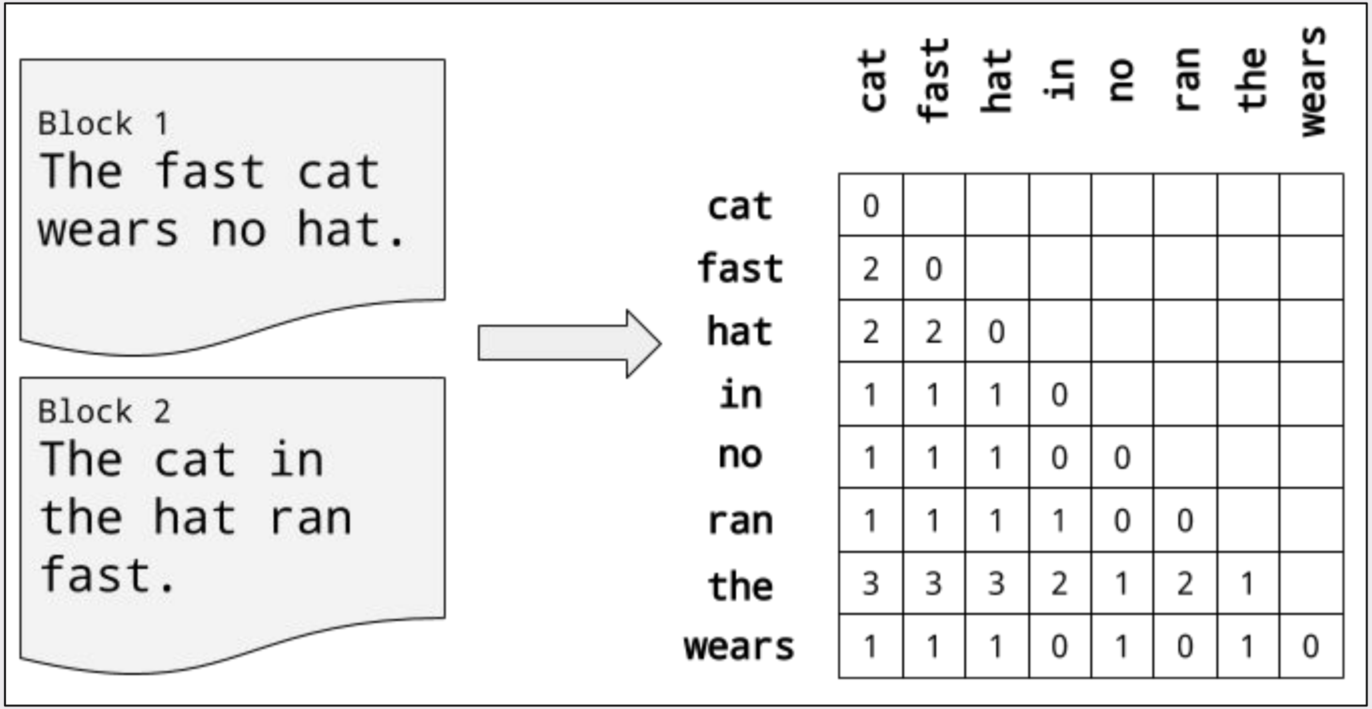
Complexity

“Colorless green ideas sleep furiously.”

Evolution

“The men replaced the battery on the hill.”

Natural Language Processing requires flexibility, which generally comes from machine learning.



Block 1
The fast cat wears no hat.

Block 2
The cat in the hat ran fast.

	cat	fast	hat	in	no	ran	the	wears
cat	0							
fast	2	0						
hat	2	2	0					
in	1	1	1	0				
no	1	1	1	0	0			
ran	1	1	1	1	0	0		
the	3	3	3	2	1	2	1	
wears	1	1	1	0	1	0	1	0

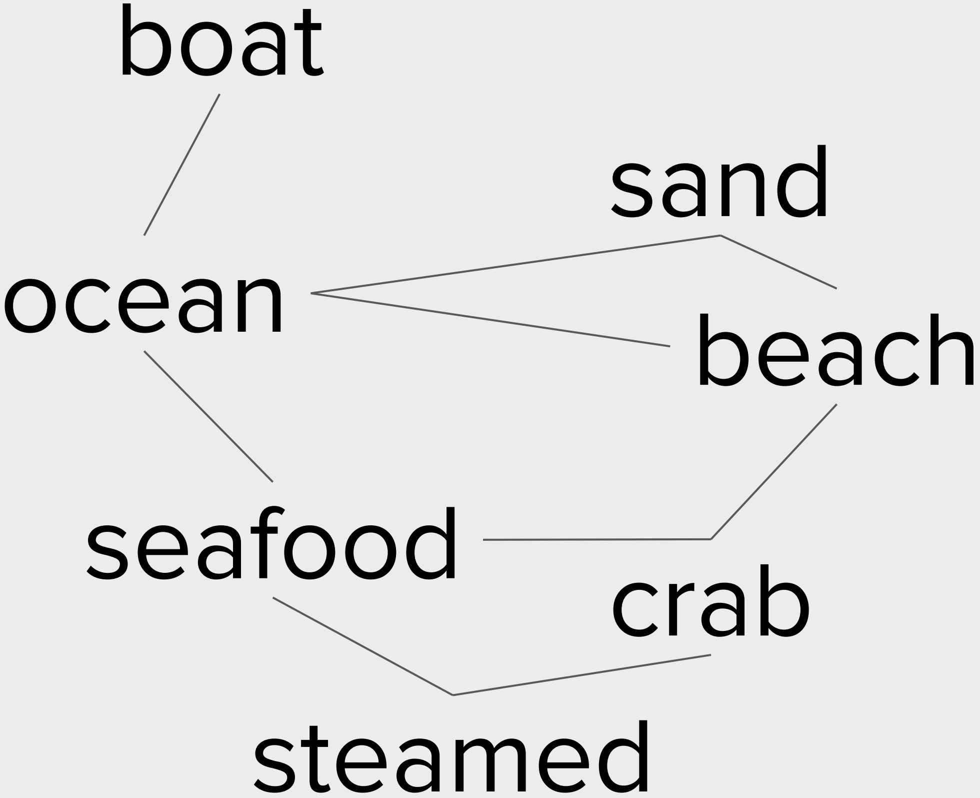
The most basic learning model is a language model.

“There was a ton of traffic on the beltway so I was ________.”

“At beach we watched the ________.”

“Watch out for that ________!”

Basic intuition: language is predictable



boat

sand

ocean

beach

seafood

crab

steamed

Models create relationships between tokens, but without meaning

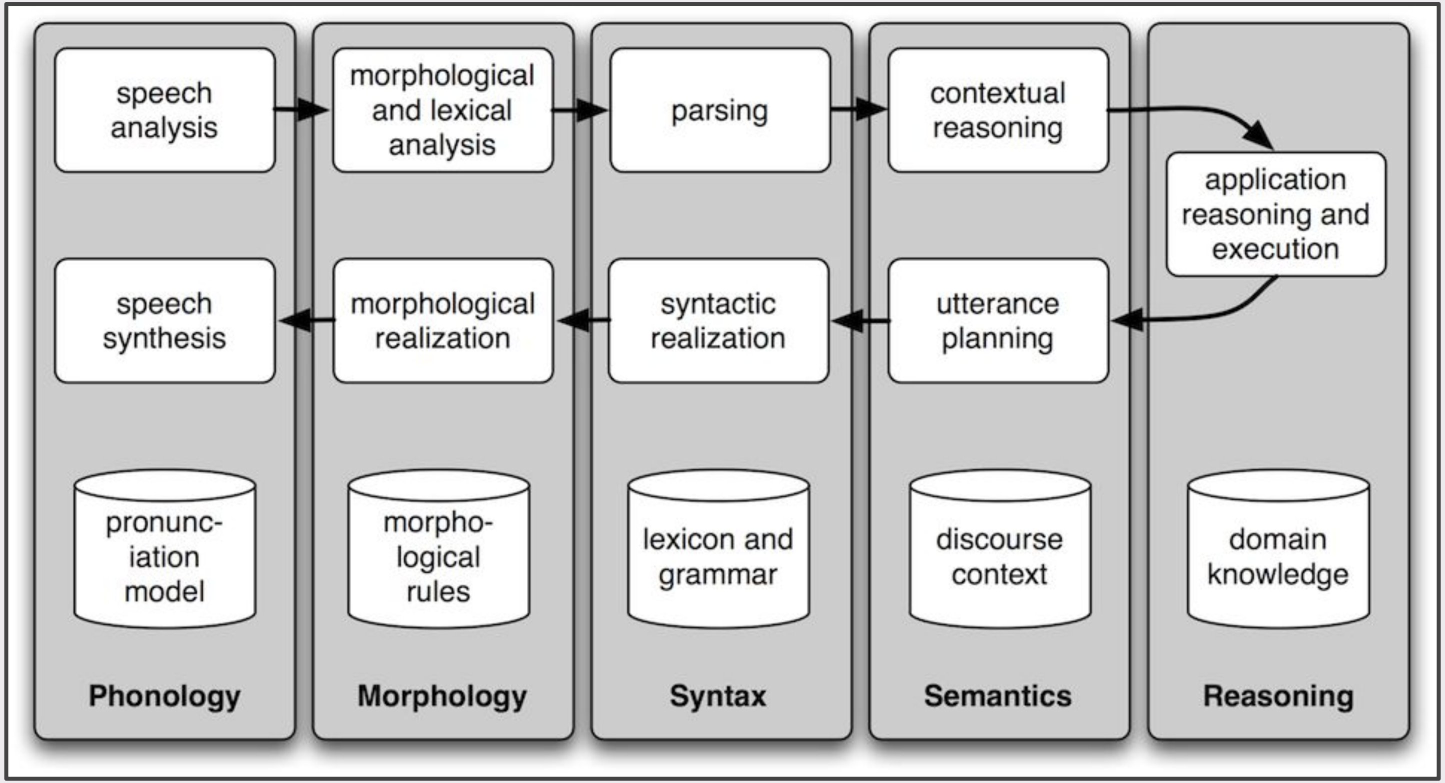
Tokens vs. Words

- Substrings
- Only structural
- Data

- Objects
- Contains a “sense”
- Meaning

```
“bearing”
“shouldn’t”
```

```
to bear.verb-1
should.auxverb-3
not.adverb-1
```



The NLP Pipeline (Language Generation)

Morphological Analysis

The study of the forms of things, words in particular.

Consider pluralization for English:

- Orthographic Rules: puppy → puppies
- Morphological Rules: goose → geese or fish

Major parsing tasks:

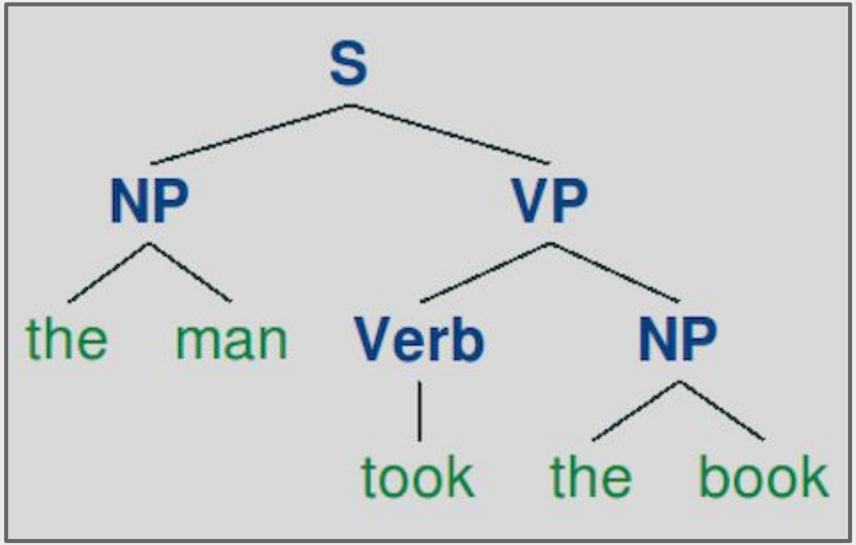
Tokenization, stemming, lemmatization and part of speech tagging.

Syntactic Analysis

The study of the rules for the formation of sentences.

Major tasks:

chunking, parsing, feature parsing, grammars



Semantic Analysis

The study of meaning.

- I see what I eat.
- I eat what I see.
- He poached salmon.

Major Tasks

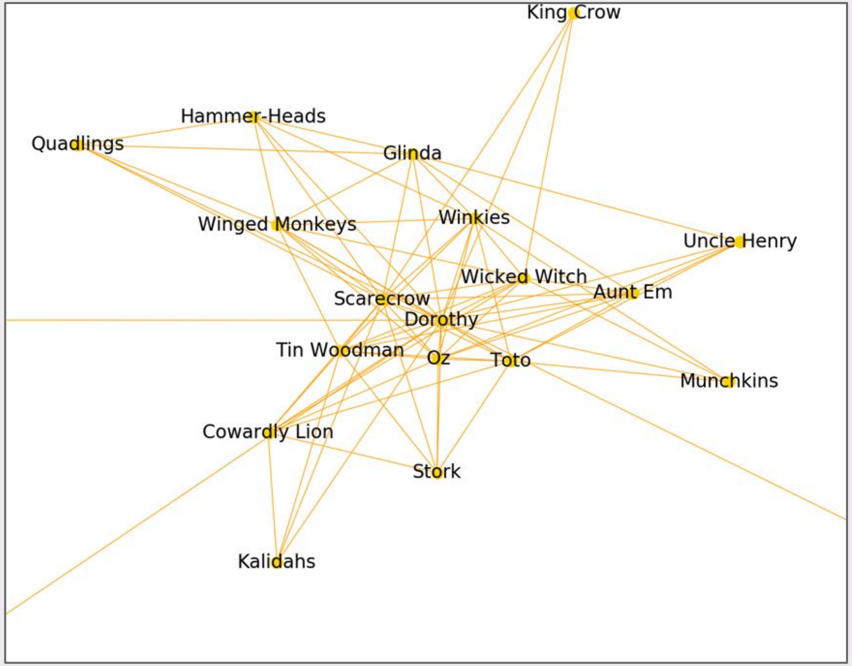
Frame extraction, Meaning Representations

Graph Analysis

Study the dynamics of complex relationships described or extracted from text.

Major tasks:

Named entity recognition, keyphrase extraction, relationship identification

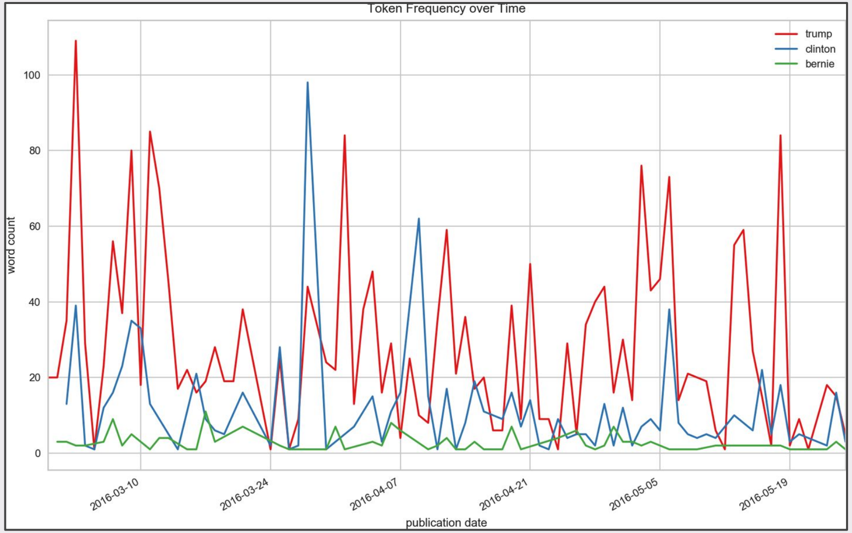


Time Series Analysis

Study trends and topics that exist in text over time.

Major tasks:

Record linkage, canonicalization, topic modelling

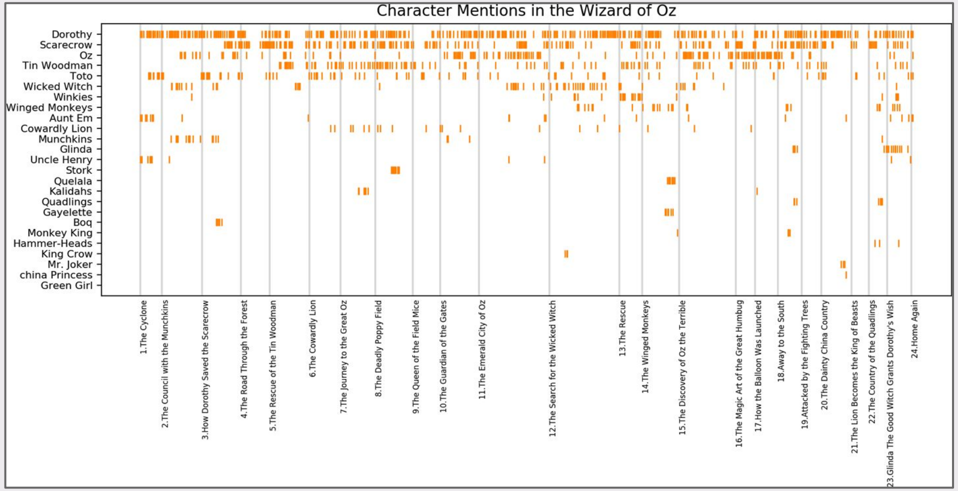


Text Visualization

Often an application in their own right, text visualization can serve as a high level summary of large amounts of information.

Major tasks:

Summarization, Visual Encoding, Relevance

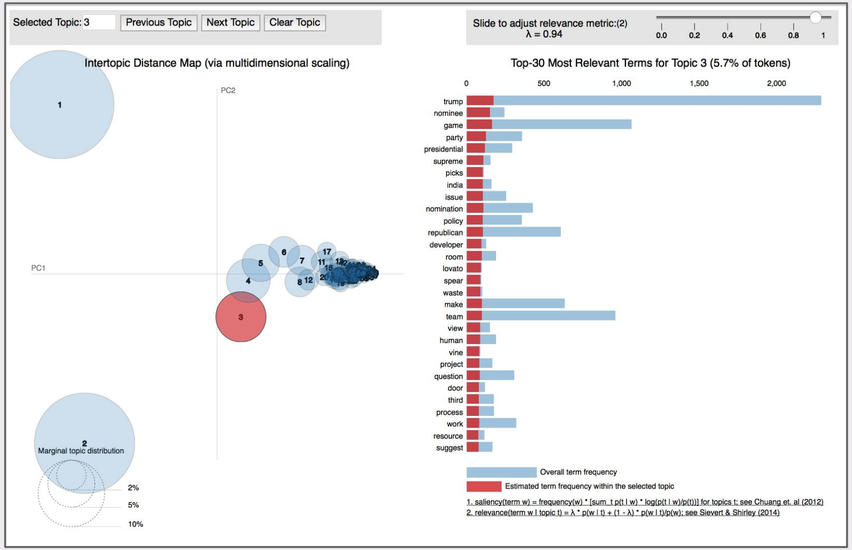


Classification and Clustering Analysis

Larger corpora of text power specific ML models to find predictive relationships between tokens.

Major tasks:

Topic Models, Sentiment Analysis



Selected Topic: 3

Slide to adjust relevance metric:(2)
λ = 0.94

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)